Advanced Regression: 1b Linear and generalised linear models (Part II)

Garyfallos Konstantinoudis

Epidemiology and Biostatistics, Imperial College London

21st February 2023

Generalised linear model (GLM)

- Linear models can only model a quantitative outcome.
- Quantitative outcomes are defined as a real number, taking possible values from $-\inf$ to $+\inf$.
- Many important data types can by definition not be modelled using a linear model:
  - Dichotomous or binary $\rightarrow$ only takes two values, 0 or 1
  - Counts $\rightarrow$ only positive integers (0,1,2,3,...)

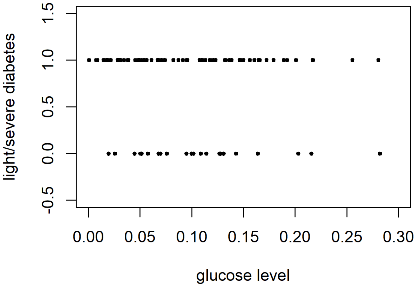
Generalised linear model (GLM)

- Flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

Binary outcome and logistic regression

Example: Case-control study

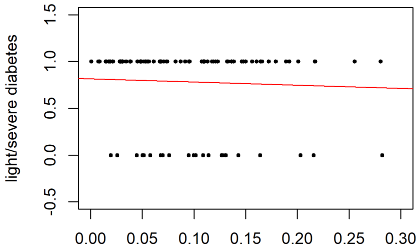
$$y_i = \begin{cases} 1 \text{ if subject } i \text{ is a case} \\ 0 \text{ if subject } i \text{ is a control} \end{cases}$$



Binary outcome and logistic regression

$$y = \underbrace{\alpha + \beta x}_{\text{Linear predictor}} + \epsilon$$

- Linear predictor $\eta = \alpha + \beta x$ is defined from $-\inf$ to $+\inf$.
- But y can only take values 0 or 1 $\rightarrow$ The linear regression fit will not match the data well.



1. Key idea:

- Instead of modelling the outcome ($y = 0$ or $y = 1$) directly, logistic regression models the probability for $y = 1$ denotes as

$$P(y = 1 \mid x)$$

Notes on probabilities for binary data:

- Probabilities can take values from 0 to 1
- Probabilities are symmetric

$$P(y = 1 \mid x) = 1 - P(y = 0 \mid x)$$

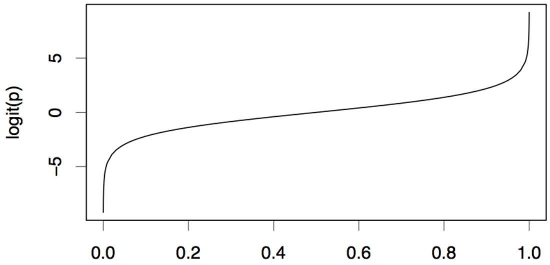
Logistic function

2. Key idea:

- Transform the linear predictor $\eta = \alpha + \beta x_i$ (quantitative, can take values from $-\inf$ to $-\inf$) to lie in the Interval [0,1], which is valid for probabilities.

- This can be achieved using the logit function:

$$\mathrm{logit(p)} = \log(\mathrm{p}/(1-\mathrm{p}))$$



Logistic regression

Logistic regression

$$\mathrm{logit}(\mathrm{P}(\mathrm{y}=1 \mid \mathrm{x})) = \log(\mathrm{P}(\mathrm{y}=1 \mid \mathrm{x})/(1-\mathrm{P}(\mathrm{y}=1 \mid \mathrm{x}))) = \alpha+\beta \mathrm{x}$$

- **Interpretation**: The regression coefficient β in logistic regression represents the log odds ratio between $y=0$ and $y=1$.
- **Estimation**: Maximum likelihood.

Technical details

- Many important outcome types can be accommodated by GLMs.
- Each of these distributions has a location parameter, e.g. μ for the Gaussian, p for the Bernoulli and Binomial.
- The natural link function between the location parameter and the linear predictor can be derived from the mathematical form of the distribution.

Response	Distribution	$E(y)$	Link (g)
Continuous	Gaussian	μ	identity
Dichotomous	Bernoulli	p	logit
Counts	Binomial	p	logit
Counts	Poisson	λ	log

Technical details: GLM

The GLM consists of three elements:

1. A probability distribution from the **exponential family**. Note: Only distributions that can be formulated as an exponential family can be modelled as GLM.
2. A linear predictor $\eta = x\beta$.
3. A link function g such that $E(y) = \mu = g^{-1}(\eta)$.

Technical details: Exponential families

An exponential family is a set of probability distributions of the following form

$$f_x(x \mid \theta) = h(x)\exp\{\eta(\theta) \times T(x) - A(\theta)\}$$

where

- θ is our parameter of interest
- $T(x)$ is a sufficient statistic.
- $\eta(\theta)$ is the natural parameter or link function.

Gaussian distribution as exponential distribution

Gaussian distribution with unknown μ, but known σ

$$
\begin{aligned}
f_\sigma(x \mid \mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{(x-\mu)^2}{2\sigma^2}\} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\}
\end{aligned}
$$

- $\theta = \mu$
- $h(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-\frac{x^2}{2\sigma^2}\}$
- $T(x) = \frac{x}{\sigma}$
- $\eta(\mu) = \frac{\mu}{\sigma}$
- $A(\mu) = \frac{\mu^2}{2\sigma^2}$

Logistic regression and binary outcomes

Binomial distribution with known number of trials n, but unknown probability p

$$\begin{aligned} f(x \mid p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} \exp\{x \log(\frac{p}{1-p}) + n \log(1-p)\} \end{aligned}$$

- $\theta = p$
- $h(x) = \binom{n}{x}$
- $T(x) = x$
- $\eta(p) = \log(\frac{p}{1-p})$
- $A(p) = -n \log(1-p)$

Logistic regression and binary outcomes

Formulate model: Three elements

1. Error distribution for response variable
2. Linear predictor
3. Link function

The three elements of the logistic regression model are

1. The Bernoulli probability distribution modelling the data: $\mathbb{P}(y_i = 1 \mid x_i) = p_i$
2. The linear predictor: $\alpha + \sum_{j=1}^{p} \beta_j x_{ij}$
3. The link function g associating the mean of y, $\mathbb{P}(y_i = 1 \mid x_i)$ to the linear predictor: here the link is the logistic link as we set $g(\mathbb{P}(y_i = 1 \mid x_i) = \mathrm{logit}(\mathrm{p_i}) = \beta_0 + \sum_{\mathrm{j=1}}^{\mathrm{p}} \beta_{\mathrm{j}} \mathrm{x_{ij}}$

glm(): in R

- GLMs can be called in R just as linear models.
  `glm(y_binary ~ age+sex+bmi+map+ltg, data = x, family=binomial)`

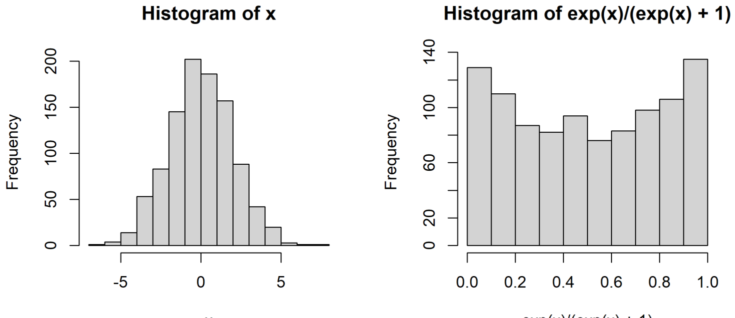
```
> summary(glm_out)

Call:
glm(formula = y_binary ~ age + sex + bmi + map + ltg, family = binomial,
    data = x)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9203  -0.5727  -0.2611   0.3643   2.9926

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6505     0.1753  -9.417  < 2e-16 ***
age          -1.6660     3.3488  -0.497    0.619
sex          -2.2971     3.0168  -0.761    0.446
bmi          21.2383     3.5556   5.973 2.33e-09 ***
map          13.3619     3.3562   3.981 6.85e-05 ***
ltg          22.2722     3.6066   6.175 6.60e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 518.87  on 441  degrees of freedom
```

glm(): in R

- Different types of exponential families can be called using the `family` option:
  - binomial(link = ‘logit’)
  - gaussian(link = ‘identity’)
  - Gamma(link = ‘inverse’)
  - inverse.gaussian(link = ‘$1/\mu^2$’)
  - poisson(link = ‘log’)
- There are similar return values as for the lm function:
  - coefficients
  - residuals
  - fitted.values
  - linear.predictors: the linear fit on link scale

Making predictions

1. Train the prediction rule.
```
glm_predict = glm(ybin_train ~ glu, data = x_train, family=binomial)
```
2. Derive predictions on the linear scale for the new data `x_tnew`.
```
eta = predict.glm(glm_predict,x_new)
```
3. Using the inverse logit transform to probabilities.



Take away: Generalised linear models

The model formulation in GLMs consists of three elements:

1. Error distribution for response variable
2. Linear predictor
3. Link function

Most common data types can be modelled using GLMs

- Continuous $\rightarrow$ Gaussian distribution
- Dichotomous or binary $\rightarrow$ Bernoulli distribution
- Counts $\rightarrow$ Poisson or Binomial (with known number of trials) distribution